Bias Considerations for Minimum Subspace Noise Tracking

Mahdi Triki and Kees Janse
Philips Research Laboratories, Eindhoven, The Netherlands
{mahdi.triki,kees.janse}@philips.com

Abstract
Speech enhancement schemes rely generally on the knowledge of the noise power spectral density. The estimation of these statistics is particularly a critical issue and a challenging problem under non-stationary noise conditions. With this respect, subspace based approaches have shown to allow for reduced estimation delay and perform a good tracking vs. final misadjustment tradeoff [2, 3]. One key attribute for noise floor tracking is the estimation bias: an overestimation leads to over-suppression and to more speech distortion; while an underestimation leads to a high level of residual noise. The present paper investigates the bias of the subspace-based scheme, and particularly the robustness of the bias compensation factor to the desired speaker characteristics and the input SNR.

Index Terms: power spectral density tracking; non-stationary noise; subspace methods; estimation bias; oversubtraction / undersubtraction.

1. Introduction
Speech enhancement aims at improving the performance of audio communication in a noisy environment. Several practical methods have already been proposed. Among them, the class of frequency domain methods has been relatively successful due to their implementation simplicity and their capability of handling noise non-stationarity to some extent. These schemes recover the clean signal by applying a gain filter. The design of these filters relies on the knowledge of the speech and noise signal statistics. In practice however, these statistics are not explicitly available and should be estimated. The accuracy of the overall enhancement approach critically depends on the estimation quality of the unknown statistics, and particularly the estimation bias: an overestimation of the spectral noise variance leads to over-suppression and to more speech distortion; while an underestimation leads to a high level of residual noise.

Joint clean speech and noise Power Spectral Density (PSD) estimation is an underdetermined problem. In fact using a unique observation, we aim at tracking both the clean speech and noise statistics. A classic trick to overcome the underdeterminacy problem is to exploit speech pauses. The key observation is that the speech signal is not always present, and the received signal during pause frames has almost no effect on the minimization results. The effective search memory is then $M' = M - M_\text{\#}$, where $M_\text{\#}$ denotes the number of speech frames, and $M$ the number of frames. Generally, the bias compensation factor depends on extrinsic variabilities, such as the desired speaker characteristics and the input SNR. However, these parameters are rarely available, and bias robustness is a critical issue for practical use. In [1], the dependence on the extrinsic variabilities was considered through the estimation and tracking of the inverse normalized variance $Q_{\text{eq}}$ (also called 'equivalent degree of freedom'). In [2], the authors have expressed this dependence as a function of the model order (signal subspace dimension). The MSNT scheme merges the consistency of the subspace structure and the flexibility of minimum statistics tracking.

Noise tracking schemes provide generally biased estimations. The noise level has to be corrected using a pre-identified bias compensation factor $B$. For instance, MS bias originates from the fact that the minimum value of a set of random variables is smaller than their mean, i.e.,

$$E\{\min(.)\} \leq \min(E\{\}.\}$$

and depends therefore on the search memory $M$. However, if the speech signal was present in $M' = M - M_\text{\#}$, these frames have almost no effect on the minimization results. The effective search memory is then $M' = M - M_\text{\#}$, and the appropriate bias compensation is rather $B = M - M_\text{\#}$, that depends on the speaker characteristics (via $M_\text{\#}$). Generally, the bias compensation factor depends on extrinsic variabilities, such as the desired speaker characteristics and the input SNR. However, these parameters are rarely available, and bias robustness is a critical issue for practical use. In [1], the dependence on the extrinsic variabilities was considered through the estimation and tracking of the inverse normalized variance $Q_{\text{eq}}$ (also called 'equivalent degree of freedom'). In [2], the authors have expressed this dependence as a function of the model order (signal subspace dimension). The MSNT scheme merges the same search approach with MS. It is thus subject to the same dependency on the extrinsic variabilities. Hereafter, we will investigate the bias of the MSNT scheme, particularly, the robustness of the bias compensation factor to the desired speaker characteristics and the input SNR.

The remainder of this paper is organized as follows. After a brief introduction to subspace-based noise tracking techniques in Section 2, we will investigate the MSNT bias estimation and robustness in Section 3. Experimental results are presented in Section 4. Finally, a discussion and concluding remarks are proposed in Section 5.

2. Subspace Decomposition for Speech Signal in DFT Domain
Classic noise floor estimation (based on voice activity detection or minimum statistics) hinges on the assumption that a speech signal is not constantly present. The received signal during pause frames is used to update the noise PSD estimate. Herein, we exploit further speech signal structures in order to get low rank model, when the noise is full rank. Therefore, a noise subspace can be identified, and the noise PSD is updated even when speech is constantly present. In [3], we have introduced a new subspace-based scheme, called Minimum Subspace Noise Tracking (MSNT). The proposed technique exploits the speech structure without an explicit identification of the signal and noise subspaces. The tracking is performed via a local search approach. In other words, the MSNT merges the consistency of the subspace structure and the flexibility of minimum statistics tracking.

Noise tracking schemes provide generally biased estimations. The noise level has to be corrected using a pre-identified bias compensation factor $B$. For instance, MS bias originates from the fact that the minimum value of a set of random variables is smaller than their mean, i.e.,

$$E\{\min(.)\} \leq \min(E\{\}.\}$$

and depends therefore on the search memory $M$. However, if the speech signal was present in $M' = M - M_\text{\#}$, these frames have almost no effect on the minimization results. The effective search memory is then $M' = M - M_\text{\#}$, and the appropriate bias compensation is rather $B = M - M_\text{\#}$, that depends on the speaker characteristics (via $M_\text{\#}$). Generally, the bias compensation factor depends on extrinsic variabilities, such as the desired speaker characteristics and the input SNR. However, these parameters are rarely available, and bias robustness is a critical issue for practical use. In [1], the dependence on the extrinsic variabilities was considered through the estimation and tracking of the inverse normalized variance $Q_{\text{eq}}$ (also called 'equivalent degree of freedom'). In [2], the authors have expressed this dependence as a function of the model order (signal subspace dimension). The MSNT scheme merges the same search approach with MS. It is thus subject to the same dependency on the extrinsic variabilities. Hereafter, we will investigate the bias of the MSNT scheme, particularly, the robustness of the bias compensation factor to the desired speaker characteristics and the input SNR.

The remainder of this paper is organized as follows. After a brief introduction to subspace-based noise tracking techniques in Section 2, we will investigate the MSNT bias estimation and robustness in Section 3. Experimental results are presented in Section 4. Finally, a discussion and concluding remarks are proposed in Section 5.

2. Subspace Decomposition for Speech Signal in DFT Domain
Classic noise floor estimation (based on voice activity detection or minimum statistics) hinges on the assumption that a speech signal is not constantly present. The received signal during pause frames is used to update the noise PSD estimate. Herein, we exploit further speech signal structures in order to get low rank model, when the noise is full rank. Therefore, a noise subspace can be identified, and the noise PSD is updated even when speech is constantly present. In [3], we have introduced a new subspace-based scheme, called Minimum Subspace Noise Tracking (MSNT). The proposed technique exploits the speech structure without an explicit identification of the signal and noise subspaces. The tracking is performed via a local search approach. In other words, the MSNT merges the consistency of the subspace structure and the flexibility of minimum statistics tracking.

Noise tracking schemes provide generally biased estimations. The noise level has to be corrected using a pre-identified bias compensation factor $B$. For instance, MS bias originates from the fact that the minimum value of a set of random variables is smaller than their mean, i.e.,

$$E\{\min(.)\} \leq \min(E\{\}.\}$$

and depends therefore on the search memory $M$. However, if the speech signal was present in $M' = M - M_\text{\#}$, these frames have almost no effect on the minimization results. The effective search memory is then $M' = M - M_\text{\#}$, and the appropriate bias compensation is rather $B = M - M_\text{\#}$, that depends on the speaker characteristics (via $M_\text{\#}$). Generally, the bias compensation factor depends on extrinsic variabilities, such as the desired speaker characteristics and the input SNR. However, these parameters are rarely available, and bias robustness is a critical issue for practical use. In [1], the dependence on the extrinsic variabilities was considered through the estimation and tracking of the inverse normalized variance $Q_{\text{eq}}$ (also called 'equivalent degree of freedom'). In [2], the authors have expressed this dependence as a function of the model order (signal subspace dimension). The MSNT scheme merges the same search approach with MS. It is thus subject to the same dependency on the extrinsic variabilities. Hereafter, we will investigate the bias of the MSNT scheme, particularly, the robustness of the bias compensation factor to the desired speaker characteristics and the input SNR.

The remainder of this paper is organized as follows. After a brief introduction to subspace-based noise tracking techniques in Section 2, we will investigate the MSNT bias estimation and robustness in Section 3. Experimental results are presented in Section 4. Finally, a discussion and concluding remarks are proposed in Section 5.
information on noise statistics even when speech is present. We focus on the time evolution of the speech DFT coefficients.

The sampled time domain signal is divided into overlapping blocks that are windowed by a smooth function, such as a Hanning window. Each windowed block is transformed into the frequency domain using a DFT. We use $s(n, f)$ and $v(n, f)$ to denote the complex DFT coefficients of the clean speech and the noise signals, respectively. $f$ represents the frequency index and $n$ the time-frame index (figure 1).

$$n - k_1 \quad n - 1 \quad n \quad n + 1 \quad n + k_2$$

Time domain signal

- DFT
- $\gamma(n-k_1,f)$
- $\gamma(n-1,f)$
- $\gamma(n,f)$
- $\gamma(n+1,f)$
- $\gamma(n+k_2,f)$

$$\gamma(n,f) = [\gamma(n-k_1,f), \ldots, \gamma(n+k_2,f)]^T$$

$$R_y(n,f) = E[\gamma \gamma^H]$$

Figure 1: Subspace decomposition in DFT domain.

We define correlation matrices in the DFT domain (for each DFT coefficient) as shown in figure 1: we collect DFT coefficients per frequency bin $f$ that originate from the time frame $n - k_1$ up to frame $n + k_2$ and we form a vector $\gamma(n, f)$ of size $K = k_1 + k_2 + 1$. The noisy speech correlation matrix (at the frequency bin $f$ and the time frame $n$) is:

$$R_y(n, f) = E[\gamma(n, f)\gamma^H(n, f)] .$$

Assuming an additive noise model, we split the noisy speech correlation matrix $R_y(n, f)$ into:

$$R_y(n, f) = R_s(n, f) + R_w(n, f).$$

We assume that the matrix $R_s(n, f)$ is rank limited. This assumption was experimentally tested and validated in [3, 2]. A relation can also be established between the rank-limited assumption and standard audio representations such as damped sinusoidal [4] and amplitude modulation [5] models.

If we assume the noise DFT correlation matrix spherical, i.e., $R_w(n, f) = \sigma_w^2(n, f)I_K$ ($I_K$ is the $K \times K$ dimensional identity matrix), the eigen decomposition of the received signal covariance matrix $R_y$ can be expressed as:

$$R_y(n, f) = U(n, f) \Lambda_s(n, f) + \sigma_w^2(n, f)I_K \quad U^H(n, f)$$

where $\Lambda_s = diag(\lambda_1, \ldots, \lambda_s, Q, 0, \ldots, 0)$ is a diagonal matrix containing the eigenvalues of $R_s$, and $U$ is an unitary matrix containing the corresponding eigenvectors. $Q$ denotes the rank of the speech correlation matrix ($Q < K$ due to the rank-limited assumption). Therefore, the noise PSD could be updated by identifying the noise subspace (estimating the model order $Q$) and averaging the $K - Q$ smallest eigenvalues [2]. Hereafter, we will refer to this scheme as Subspace Noise Tracking (SNT) technique.

Due to the bursty nature of the speech signal and the continuous distribution of the speech eigenvalues, the rank of the speech covariance matrix $Q$ is difficult to estimate and speech/noise subspaces are hard to separate. In [3], we have proposed a scheme exploiting the subspace structure without an explicit model order selection. Only the minimum eigenvalue $\lambda_{s,min}(n, f)$ is considered to update the noise PSD. Indeed, it has been observed that the $\lambda_{s,min}$ is less depending on speech signal and provides more consistent information on the noise PSD. Moreover, the estimation and the adaptation of $\lambda_{s,min}$ could be implemented in a computationally efficient way (no need to perform complete eigenvalue decomposition).

The proposed scheme assumes also that speech DFT coefficients live often in a low dimension subspace. In other words, for a given frequency bin and using a sufficiently long memory $M$, at least one speech DFT covariance matrix $R_s(n, f)$ is rank limited. Exploiting the observation that the minimum power level reaches the noise level, noise tracking is performed via local search approach, i.e.,

$$\hat{\sigma}_w^2(n, f) = \frac{1}{B} \min_{i=0:M-1} \{\lambda_{s,min}(n - i, f)\}$$

$B$ denotes a bias compensation factor (discussed in the following sections). Compared to the MS approach, the MSNT further exploits the structure of the speech signal time variation (not only in terms of presence or absence). It has been observed that the MSNT does not need a large search memory $M$ to perform an acceptable steady state performance, and provides better noise tracking [3].

3. Bias Compensation for Minimum Subspace Noise Tracking

The MSNT method rests on two observations:

(A1) speech and noise are statistically independent.

(A2) speech DFT coefficients live often in a low dimension subspace.

The MSNT estimation approach tracks the minimum of a short-time PSD estimate $\lambda_{s,min}$ within a finite window of length $M$. Since for non-trivial densities the minimum value of a set of random variables is smaller than their mean, the MSNT estimate is necessarily biased. This bias (that originates from minimum search based tracking) depends mostly on the search memory $M$, and may be corrected using a multiplicative bias compensation factor [1]. The focus of this section is to derive the bias compensation factor and analyze the bias compensation robustness to extrinsic variabilities.

The bias can be analytically computed only if the random sequence $\{\lambda_{s,min}(n, f)\}_n$ is independent, identically distributed (i.i.d). As the covariance matrix $R_y$ is computed using a sliding processing, this assumption is clearly not valid. We therefore develop an approximate estimation of the bias compensation factor $B(M)$. We make use of a training procedure based on speech data degraded with white noise with a known variance $\sigma_w^2(n, f) = \sigma_w^2(n, f)$. At each time-frequency bin $(n, f)$ the local bias is computed:

$$B_M(n, f) = \frac{1}{\#(n, f)} \min_{i=0:M-1} \{\lambda_{s,min}(n - i, f)\}$$

The global bias compensation factor is derived by averaging the local bias over time-frequency axis, i.e.,

$$B(M) = \frac{1}{\#(n, f)} \sum_{(n, f)} B_M(n, f)$$

where $\#(n, f)$ denotes the total number of time-frequency frames used for bias learning. Note that if the tracking memory
$M$ is kept constant, the bias is compensated using a constant multiplicative factor. In several applications, such as for noise whitening pre-processing, noise PSD need only to be estimated up to a multiplicative factor. In such a case, bias compensation is not required (contrary to the SNT bias compensation, dependent on the model order and variant in time-frequency domain).

Strictly speaking, the bias compensation factors depend both on the desired speaker and the input SNR. However, these parameters are hardly available. Bias learning robustness is critical for practical use. First, we investigate the bias learning robustness with respect to the speaker characteristics. The SNT and MSNT tracking bias are learned as described in [2] and (6) using four different speakers (2 males and 2 females), respectively. The input SNR is 5 dB. Fig. 2 plots the SNT tracking bias function of the model order $Q$.

Figure 2: Robustness of the SNT tracking bias to speaker characteristics.

One can remark that the bias shape depends significantly on the speaker gender. Therefore, it constitutes a major drawback in case the desired speaker (or at least the gender) is unknown.

Next, we consider the robustness of the tracking bias with respect to the input SNR. The bias is learned by combining male and female speakers under three input SNR conditions (0dB, 5dB, 10dB). Fig. 4 and 5 plot respectively the SNT and MSNT tracking bias function of the input SNR.

Figure 4: Robustness of the SNT tracking bias to input SNR.

One may observe that bias compensation factors decrease with increased input SNR and that MSNT bias robustness improves with increased search memory. Subjective tests show that, contrary to speaker characteristics, the input SNR has no major effect on the overall enhancement accuracy: most of the errors are corrected by the use of noise over/under-subtraction [7, 8].
4. Experimental Results

We evaluate the accuracy of the different noise tracking schemes under non-stationary noise conditions. The speech signal originates from the NOIZEUS database [6] (sampled at 8 kHz). A non-stationary noise is synthetically scaled (at an input SNR of 5 dB), then added to the clean speech signal. We consider two non-stationary noise sources: originated respectively from a passing car and a passing train. The noisy signal is segmented using a square root Hanning window (length= 256, overlap=87.5%). The dimension of the DFT covariance matrix is $K = 7$. The DFT covariance matrix was estimated over the time frame $[n-n_1 : n+n_2]$ (we have set $n1 = n2 = 7$). The search memory of the MSNT scheme was set to $M = 60$. Both SNT and MSNT bias compensation factors are trained combining male and female speakers (different from those used for the experiments). To illustrate the noise tracking accuracy, we plot the true and estimated noise level (averaged over frequencies) function of the frame index in figures 6 and 7.

Curves illustrate that SNT locally overestimates the noise PSD. Subjective tests demonstrate that such artifact spoils weak speech sounds and therefore reduces intelligibility. On the other hand, MSNT estimate is less subject to overestimation, because only minimum eigenvalue is used for tracking (limiting the effect of speech leakage). The figures show also that although MSNT shares with MS the bias source (local search based update), MSNT performs better tracking and less underestimates noise PSD. Indeed, MSNT tracks the noise PSD in frames where no-speech is present in some directions, where MS tracks the noise PSD in frames where no-speech is present in all directions, which are (intuitively) more frequent and seems to be more uniformly distributed.

5. Concluding Remarks

In the present paper, we have investigated the estimation bias of subspace-based method. Having an unbiased noise estimate is crucial for enhancement applications: an overestimation leads to over-suppression and to more signal distortion; while an underestimation leads to a high level of residual noise. Generally, the bias is corrected using a (formerly computed) bias compensation factor. The robustness of the compensation factor to extrinsic variabilities is then of utmost importance. With this respect, it has been shown that, compared to SNT, MSNT offers superior and homogeneous bias robustness (particularly w.r.t. speaker characteristics variabilities). Simulation shows that MSNT performs better non-stationary noise floor tracking (compared to both MS and SNT), leading to a better quality vs. intelligibility tradeoff.

6. References


